A Belief-Based Approach to Measuring Message Acceptability

Célia da Costa Pereira, Andrea G. B. Tettamanzi, and Serena Villata

Université Côte d'Azur, CNRS, Inria, I3S, France
{celia.pereira,andrea.tettamanzi,serena.villata}@unice.fr

Abstract. We propose a formal framework to support belief revision based on a cognitive model of credibility and trust. In this framework, the acceptance of information coming from a source depends on *(i)* the agent's goals and beliefs about the source's goals, *(ii)* the credibility, for the agent, of incoming information, and *(iii)* the agent's beliefs about the context in which it operates. This makes it possible to approach belief revision in a setting where new incoming information is associated with an acceptance degree. In particular, such degree may be used as input weight for any possibilistic conditioning operator with uncertain input (i.e., weighted belief revision operator).

1 Introduction

Fulfilling its goals is an important concern for an agent. The agent's perceptions about its environment highly influences this process. Such perceptions dynamically enrich the agent's beliefs, namely thanks to new more or less credible/trusted information. According to the principle of *primacy update*, in belief revision new information is generally accepted. However, as pointed out by several authors [10, 11, 13], in real-world situations it is often the case that new information is not fully considered or simply not accepted due to an insufficient amount of plausibility [5]. The extent to which new information will be accepted (i.e., really considered) by the agent directly depends on these credibility and trust values. A key factor for the agent's success in fulfilling its goals is then its ability to compute both the *credibility* of new information and its *trust* in the source providing information. Recently, Adali [1] has proposed to define information trust as a computational concept whose value depends on the trustworthiness of the information source,[1] and on the credibility of the information content. Adali's approach also agrees with the one proposed by Sparks [26]. However, to the best of our knowledge, a formal framework for measuring the acceptability of a message which takes into account the agent's goals and the agent's beliefs about the source's goals, the credibility of the message with respect to its content and with respect to both the agent's competence and the source's competence, and the agent's beliefs about the source's nature (malicious or not) is still missing.

The research question of how to provide such a framework breaks down into the following subquestions:

[1] In the rest of the paper we will also refer to the "trustworthiness of an information source" for an agent as the agent's "trust in a source".

– How do we take the *nature* of the information source into account?
– How do we take the agent's and source's goals into account?
– How do we measure the credibility of information based on the agent's and the source's competences?

We answer these questions by proposing a possibilistic model whereby the cognitive notions of trust and credibility can be formalized and the acceptability of the pieces of information can be computed. Our framework makes it possible to:

– represent the fact that the agent's beliefs may include the "nature" of a source, which may be categorized as malicious, rational, etc.—each evaluation of a component of trust should consider this fact;
– somehow measure the source's willingness to cooperate thanks to the agent's perceptions about the source's goals—a source sharing my goals should (implicitly or explicitly) act/help for the achievement of these goals, unless, perhaps, it is not rational;
– compute the credibility of information coming from a given source with respect to:
  • the agent's competence—the agent may be able to evaluate the information content regardless of how trustworthy it considers the information source;
  • the source's competence—we suppose that (1) each piece of information belongs to a domain of competence and (2) the agent has beliefs about the domains of competence of the sources. It is then possible to evaluate the credibility of such a new piece of information with respect to the source's competence;

Our approach is cast within the framework of possibility theory in order to cover cases when not enough data to compute probabilities are available.

The paper is organized as follows: first, we present some related work and we compare them to the proposed approach. Then, we provide some basic notions of possibility theory, upon which our model is built. Our proposal is put forth in Section 4, and its formal properties are discussed. Conclusions end the paper.

2 Related Work

In multi-agent systems, representing and making possible the evaluation of the credibility associated with a piece of information is important especially when the agents have their own beliefs and can obtain new information from other sources. In this case, assessing to which extent such new information should be integrated with the agents' beliefs depends on its credibility and on the trustworthiness of its source. Tamargo *et al.* [27] address this problem in a collaborative multi-agent system in which agents can receive new information from informant agents through communication. The authors consider a credibility order among the informant agents. A belief is then revised when new contradictory incoming information arrives from an informant that is highly credible. Unlike in Tamargo's approach, where credibility is associated to agents, here we propose to associate a (computed) credibility degree to the new piece of information.

Krümpelmann *et al.* [14] propose to attach an agent identifier to each piece of information, representing the credibility of the transferred information. But still, credibility is associated to an agent. Besides, while our value of credibility together with

the trust value will determine the extent to which the new piece of information will be accepted, in the above-mentioned approaches the aim of the credibility order is to help in the choice of which, of the old belief and the new piece of information, will be adopted/maintained.

On the other hand, there exist several works about trust in the literature and in different disciplines [19,20,26,31]. Among the numerous and interesting contributions by Falcone *et al.*, we can underline [12], in which the authors claim that an agent's decision about trusting an information source or not depends on the agent's representation of the source's nature. The principle according to which "only an agent endowed with goals and beliefs can trust another agent" has been pointed out by Castelfranchi and Falcone [7]. Trust is thus considered as a matter of utility and a context-sensitive concept. All the above proposals lead us to argue that trust is a multidimensional concept. Sabater *et al.* [23] share this point of view. Indeed, they proposed a model which deals with three dimensions of trust or reputation. The first dimension is based on an agent's own experiences. The second dimension is based on third-party information obtained thanks to the agent's social relationships, and the third dimension, also called the ontological dimension, helps to transfer trust information between related contexts. Sierra and Debenham [25] propose a trust-based decision model to be used in the context of negotiation. They propose a probabilistic method to represent and define trust as depending on the information gain caused by a piece of evidence—the more information an agent has about an event, the smaller its (probabilistic) uncertainty about that event. Probability theory is also used by Teacy *et al.* [28] to represent trust by taking past interactions with other agents into account while possibility theory is used in [2] for proposing an interval-based representations of trust and distrust based on past performances by considering the fact that data are not necessarily numerous in practice.

3 Background

In this section, we provide basic notions of possibility theory and define how beliefs and goals are formalized in our framework to model cognitive agents. Finally, we propose a way to associate the information content of a message to domains of competence, by adopting implication in logical Information Retrieval models.

3.1 Language and Interpretations

A classical propositional language may be used to represent information for manipulation by a cognitive agent.

Definition 1. *(Language) Let* Prop *be a* finite *set of atomic propositions and let* $\mathcal{L}$ *be the propositional language such that* $\mathrm{Prop} \cup \{\top, \bot\} \subseteq \mathcal{L}$*, and,* $\forall \phi, \psi \in \mathcal{L}$*,* $\neg\phi \in \mathcal{L}$*,* $\phi \wedge \psi \in \mathcal{L}$*,* $\phi \vee \psi \in \mathcal{L}$*.*

As usual, one may define additional logical connectives and consider them as useful shorthands for combinations of connectives of $\mathcal{L}$, e.g., $\phi \supset \psi \equiv \neg\phi \vee \psi$. We will denote by $\Omega = \{0,1\}^{\mathrm{Prop}}$ the set of all interpretations on Prop. An interpretation $\mathcal{I} \in \Omega$ is a function $\mathcal{I} : \mathrm{Prop} \to \{0,1\}$ assigning a truth value $p^{\mathcal{I}}$ to every atomic proposition

$p \in \mathrm{Prop}$ and, by extension, a truth value $\phi^{\mathcal{I}}$ to all formulas $\phi \in \mathcal{L}$.[2] We will denote by $[\phi]$ the set of all models of ϕ, $[\phi] = \{\mathcal{I} : \mathcal{I} \models \phi\}$.

3.2 Possibility Theory

Fuzzy sets [32] are sets whose elements have degrees of membership in $[0, 1]$. Possibility theory is a mathematical theory of uncertainty that relies upon fuzzy set theory, in that the (fuzzy) set of possible values for a variable of interest is used to describe the uncertainty as to its precise value. At the semantic level, the membership function of such set, π, is called a *possibility distribution* and its range is $[0, 1]$. A possibility distribution can represent the beliefs of an agent: $\pi(\mathcal{I})$ represents the degree of compatibility of the interpretation $\mathcal{I}$ with the available evidence about the real world if we are representing uncertain beliefs. By convention, $\pi(\mathcal{I}) = 1$ means that it is totally possible for $\mathcal{I}$ to be the real world, $1 > \pi(\mathcal{I}) > 0$ means that $\mathcal{I}$ is only somehow possible, while $\pi(\mathcal{I}) = 0$ means that $\mathcal{I}$ is certainly not the real world.

A possibility distribution π is said to be normalized if there exists at least one interpretation $\mathcal{I}_0$ s.t. $\pi(\mathcal{I}_0) = 1$, i.e., there exists at least one possible situation which is consistent with the available knowledge.

Definition 2 (Fuzzy Measure). *Let Ω be a universe of discourse; a function $f : 2^{\Omega} \rightarrow [0, 1]$ is a* fuzzy measure *if*

1. $f(\emptyset) = 0$*;*
2. *for all $A, B \subseteq \Omega$, $A \subseteq B \Rightarrow f(A) \leq f(B)$.*

A fuzzy measure f is normalized *if $f(\Omega) = 1$.*

Definition 3 (Possibility and Necessity Measures). *A possibility distribution π induces a* possibility measure *and its dual* necessity measure*, denoted by Π and N respectively. Both measures apply to a classical set $S \subseteq \Omega$ and are defined as follows:*

$$\Pi(S) = \max_{\mathcal{I} \in S} \pi(\mathcal{I}); \tag{1}$$

$$N(S) = 1 - \Pi(\bar{S}) = \min_{\mathcal{I} \in \bar{S}} \{1 - \pi(\mathcal{I})\}. \tag{2}$$

A few properties of Π and N induced by a normalized possibility distribution on a finite universe of discourse Ω are the following. For all subsets $S \subseteq \Omega$:

1. $\Pi(A \cup B) = \max\{\Pi(A), \Pi(B)\}; \quad N(A \cap B) = \min\{N(A), N(B)\};$
2. $\Pi(A \cap B) \leq \min\{\Pi(A), \Pi(B)\}; \quad N(A \cup B) \geq \max\{N(A), N(B)\};$
3. $\Pi(\emptyset) = N(\emptyset) = 0; \quad \Pi(\Omega) = N(\Omega) = 1;$
4. $\Pi(S) = 1 - N(\bar{S})$ (duality);
5. $N(S) > 0 \Rightarrow \Pi(S) = 1; \quad \Pi(S) < 1 \Rightarrow N(S) = 0;$

In case of complete ignorance on S, $\Pi(S) = \Pi(\bar{S}) = 1$ and $N(S) = N(\bar{S}) = 0$.

[2] When $\phi^{\mathcal{I}} = 1$, i.e., $\mathcal{I}$ satisfies formula ϕ, in symbols $\mathcal{I} \models \phi$, $\mathcal{I}$ is called a model of ϕ.

3.3 Beliefs

We assume a possibilistic BDI model of agency like the one proposed in [9]. In that model, the epistemic state of an agent is represented by a normalized possibility distribution $\pi : \Omega \to [0, 1]$. The degree to which a given arbitrary formula $\phi \in \mathcal{L}$ is believed can, therefore, be calculated from it as

$$\mathbf{B}(\phi) = N([\phi]) = 1 - \max_{\mathcal{I} \not\models \phi} \{\pi(\mathcal{I})\}. \tag{3}$$

Straightforward consequences of the properties of possibility and necessity measures are that $\mathbf{B}(\phi) > 0 \Rightarrow \mathbf{B}(\neg\phi) = 0$, i.e., if the agent somehow believes ϕ then it cannot believe $\neg\phi$ at all; $\mathbf{B}(\phi \wedge \psi) = \min\{\mathbf{B}(\phi), \mathbf{B}(\psi)\}$ and $\mathbf{B}(\phi \vee \psi) \geq \max\{\mathbf{B}(\phi), \mathbf{B}(\psi)\}$. Notice that $\mathbf{B}(\top) = 1$ and $\mathbf{B}(\bot) = 0$.

The rationale for choosing possibility theory to represent beliefs is its ability to capture epistemic uncertainty. It is well known that possibility theory is suited to represent uncertainty by only using a notion of order (much easier to have with few data) between the possible outcomes. A viable alternative would be the Dempster-Shafer theory of evidence [24]; however, the use of that theory would be computationally much heavier, due to the need to maintain a probability mass assignment to every element of 2^{Ω}, as compared to a possibility assignment to every interpretation of Ω in possibility theory.

3.4 Goals

The goals of an agent may be represented as a set G of formulas from the same language $\mathcal{L}$. The meaning of saying that $\psi \in G$ is a goal for the agent is that the agent would be happy with any state of the world $\mathcal{I} \in \Omega$ such that $\mathcal{I} \models \psi$.

3.5 Domains of Competence

We propose to associate formulas to domains of competence. This is in line with what has been proposed by Paglieri et al. [18], except that they referred to arguments instead of just formulas and they defined domains based on the propositions their truth depends on. We propose a more general definition inspired by the use of implication in logical Information Retrieval models [30]. More precisely, given a domain d described by a formula χ_d (like a query in (fuzzy) set-based models of information retrieval) and a formula ϕ (like a document), we use implication to determine if ϕ is relevant to d i.e., if $\chi_d \models \phi$. The intuitive meaning of this is that incoming information is relevant to a domain if the models of the formula describing the domain are included in the models of the formula describing incoming information. However, because entailment is too rigid a relation and cannot express partial relevance [15], what we propose is in line with fuzzy set-based models in Information retrieval [29], where one resorts to a fuzzy measure of the $\chi_d \models \phi$ entailment. We define one such measure based on possibilistic conditioning [4] of ϕ by χ_d.

Definition 4. *Given language $\mathcal{L}$ and D the set of domains of competence, such that every $d \in D$ is defined by a formula $\chi_d \in \mathcal{L}$, the association between formulas and*

domains is represented by a fuzzy relation $R : \mathcal{L} \times D \to [0, 1]$ *such that, given* $\phi \in \mathcal{L}$*,* $d \in D$*, the membership degree of formula* ϕ *in domain* d *is*

$$R(\phi, d) = \begin{cases} 1, & \textit{if } \chi_d \models \phi, \\ \Pi([\phi \wedge \chi_d]), & \textit{otherwise.} \end{cases}$$

In addition, we may require that the domains D form a partition of the universe of discourse, i.e., that

$$\bigvee_{d \in D} \chi_d = \top, \quad \forall d_1, d_2 \in D, \chi_{d_1} \wedge \chi_{d_2} = \bot.$$

Proposition 1. *Let* $\phi, \psi \in \mathcal{L}$*. For all domain* $d \in D$*, if* $\phi \models \psi$*,* $R(\phi, d) \leq R(\psi, d)$*.*

Proof: Given a domain d, we may distinguish three cases:

1. $\chi_d \models \phi$; in this case, it must also be that $\chi_d \models \psi$ and, as a consequence, $R(\phi, d) = R(\psi, d) = 1$, and the thesis holds;
2. $\chi_d \not\models \phi$ and $\chi_d \models \psi$; in this case, $R(\phi, d) = \Pi([\phi \wedge \chi_d]) \leq 1$ and $R(\psi, d) = 1$, and the thesis holds;
3. $\chi_d \not\models \phi$ and $\chi_d \not\models \psi$; in this case, $R(\phi, d) = \Pi([\phi \wedge \chi_d])$ and $R(\psi, d) = \Pi([\psi \wedge \chi_d])$; now, $\phi \models \psi$ means $[\phi] \subseteq [\psi]$; therefore, $[\phi] \cap [\chi_d] \subseteq [\psi] \cap [\chi_d]$, hence $\Pi([\phi \wedge \chi_d]) \leq \Pi([\psi \wedge \chi_d])$, and the thesis holds.

□

4 A Formal Framework of Cognitive Trust

We are now ready to formalize the notion of trust, the nature of an information source, the relation between beliefs and goals, and credibility.

4.1 Trust as Belief

Some pieces of information can contribute to increase or decrease the trust that an agent has in a source, and others can contribute to increase or decrease distrust. Trust is also a matter of competences.[3] Indeed, we can have different evaluations of trust (distrust) in the same source relevant to different domains of competence.

Like in [17], we suppose that trust and distrust are not the opposite ends of a single continuum, but linked dimensions that can coexist and have different antecedents and consequences [20]. We consider the social-cognitive model of trust [7, 22], in which trust is defined as beliefs: an agent trusts a source s, in a domain d, if and only if it somehow believes that s will be able to somehow help it fulfill its goals. We will also define distrust as a belief: an agent distrusts a source s with respect to a domain of competence d if and only if it somehow believes that s might try to prevent it to reach its goals. Although in the next sections we will show how to compute trust and distrust

[3] Here, we name such a competence-based trust *credibility*.

in a source s, we should always keep in mind that trust and distrust in s are to be construed conceptually as if they were defined as follows:

$$\mathrm{trust}(s) \equiv \mathbf{B}(\text{“}s \text{ is trustworthy”}), \tag{4}$$

$$\mathrm{distrust}(s) \equiv \mathbf{B}(\text{“}s \text{ is untrustworthy”}). \tag{5}$$

Notice that proposition “s is untrustworthy” is the logical negation of “s is trustworthy”.

Some authors treat “distrust” as if it were defined as $\neg\mathbf{B}(\text{“}s \text{ is trustworthy”})$, in which case $\mathrm{distrust}(s) = 1 - \mathrm{trust}(s)$: trust is considered as the complement of distrust [31]. Here, we give distrust a stronger meaning: we distrust someone if we have valid reasons to believe he is lying, not if we do not have valid reasons to believe he is telling the truth. In other words, distrust is not the complement of trust.

A consequence of Equations 4 and 5, together with the properties of Π and N, is that trust and distrust, wrt a given domain, obey the following mutual constraints:

$$\mathrm{trust}(s) > 0 \Rightarrow \mathrm{distrust}(s) = 0, \tag{6}$$

$$\mathrm{distrust}(s) > 0 \Rightarrow \mathrm{trust}(s) = 0. \tag{7}$$

In case of total ignorance, we have that $\mathrm{trust}(s) = \mathrm{distrust}(s) = 0$. Notice that if we consider trust as the complement of distrust, we cannot represent the situations of total ignorance in which the agent does not know anything which could lead it to trust or distrust the source; $\mathrm{distrust}(s) = 1 - \mathrm{trust}(s) = 0.5$ would not mean ignorance!

4.2 The Nature of a Source

Any judgment about the competence or willingness of a source to provide useful information must be, implicitly or explicitly, based on an agent's judgment (i.e., beliefs) about the source according to past interactions as well as recommendations or the source's reputations. We will refer to such assessment as the source's *nature*.

Without any claim of exhaustiveness and just to ground our presentation on an intuitive setting, we draw inspiration from the abstract model of a human agent's social behavior proposed by Italian economist Carlo Cipolla [8] as the backdrop on which his theory of human stupidity is expounded.

According to Cipolla's model, an agent's behavior may be summarized by two coordinates:

- x the average gain (or loss) that an agent obtains as a result of his or her actions;
- y the average gain (or loss) that an agent produces to other agents or groups of agents.

As a result, agents can be plotted as points on a diagram like the one shown in Figure 1 based on their $\langle x, y\rangle$ behavior. Such a diagram divides the two-dimensional plane into four quadrants or eight sectors, corresponding to different natures of the agents.

For the sake of simplicity, let us represent an agent's position in one of the eight sectors by means of three propositional variables: r if agent s is *rational* ($x > 0$); m if agent s is *malicious* ($y < 0$); and e if $|y| < |x|$.

It is worth mentioning that this concept of *source nature* allows us to model the two kinds of beliefs, namely “willingness belief” and “persistence belief”, proposed by

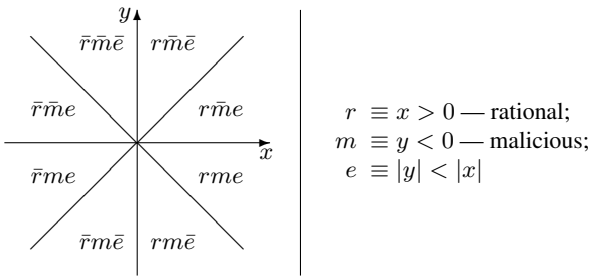

Fig. 1. The correspondence between the eight sectors of the source nature diagram and the truth assignments to the propositional variables r, m, and e.

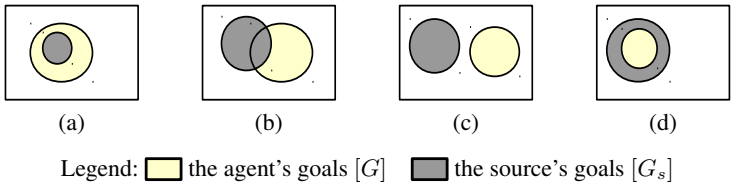

Fig. 2. A schematic illustration of the four cases of agent's goal-source'goals relationships.

Ramchurn [21] to ensure that a certain task can be delegated by an agent to another one. More precisely, the eight sectors we identify are used by the agent to decide when it needs to maintain a suspicious attitude in dealing with malicious or irrational agents.

4.3 Trust and (Shareable) Goals

We make the assumption that the agent's beliefs about the source's goals may also influence its trust in the source. Indeed, regardless of the content of information, if, for example, a source s shares the same goal g with the agent, we may suppose that s will act to fulfill g. This should, at least in case of rational sources, prevent s from taking actions that could negatively influence the satisfaction of g.

The way beliefs about the source's goals, G_s are taken into account is by comparing them with the agent's own goals, G. We distinguish four cases, which represent three possible situations. The four cases are schematically illustrated in Figure 2.

(a) $[G_s] \subseteq [G]$ or, equivalently, $G_s \models G$: if the source achieves its goals, the agent does too (*necessary help*).
(b) $[G_s] \cap [G] \neq \emptyset$ and $\overline{[G_s]} \cap \overline{[G]} \neq \emptyset$: the agent's and the source's goals are independent: the fact that either of the two achieves its goals does not necessarily imply or exclude that the other does; there is thus room for cooperation (*compatibility*).
(c) $[G_s] \cap [G] = \emptyset$: there is an overt conflict between the agent's and the source's goals (*conflict*).
(d) $[G] \subset [G_s]$ or, equivalently, $G \models G_s$: if the agent achieves its goals, the source does too, but not *vice versa* (*compatibility*).

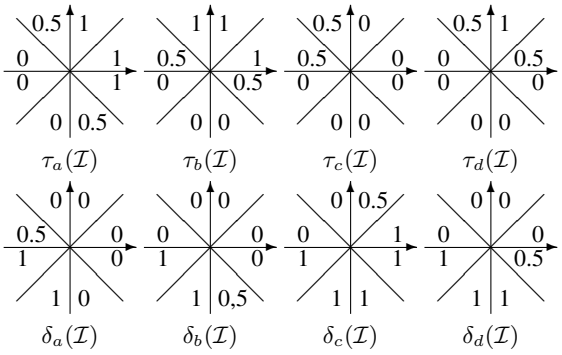

Fig. 3. A definition of functions $\tau_z(\mathcal{I})$ and $\delta_z(\mathcal{I})$, for $z \in \{a, b, c, d\}$ and $\mathcal{I} \in \{0,1\}^{\{r,m,s\}}$.

4.4 Trust in a Source

We assume that an agent has an internal reasoning mechanism allowing it to compute the trust/distrust, τ_z/δ_z, with $z \in \{(a), (b), (c), (d)\}$ (the four cases in Section 4.3). Such degrees depend on the agent's beliefs about the source's position in the nature diagram and its beliefs about the source's goals with respect to its own goals. Figure 3 shows a minimal such mechanism based on look-up tables.

Computing trust can be seen as a set of material implications. Given a source's position in the nature diagram, "if the source's goals configuration is z, then the agent will associate a trust τ_z to that source". However, the agent may not know precisely the source's position and it can just have a notion of order about which among the eight possible sectors the source could be in, some of them being more possible than others. If we consider a source s, the uncertainty is captured in our formalism trough the possibility distribution on the worlds (i.e., interpretations) $\mathcal{I}$ which are consistent with $\{0,1\}^{\{r_s,m_s,e_s\}}$. The above implication is then represented as a fuzzy implication. Among the existing definitions of fuzzy implications (see for example [16] for a survey) we adopt the Kleene-Dienes fuzzy implication. Other definitions might be used as well. The truth value of the fuzzy implication "If a source is somehow compatible with situation $\mathcal{I}$, then the agent trusts that source to degree $\tau_z(\mathcal{I})$" quantifies to what extent "the agent trusts that source to degree $\tau_z(\mathcal{I})$" is at least as true as "that source is somehow compatible with situation $\mathcal{I}$". Let us recall that we consider eight possible positions and that we have a possibility distribution on these positions. We have then eight fuzzy implications with their respective truth values. Therefore, for each goal configuration z, we define the trust and distrust that the agent has in source s as follows:[4]

$$\mathrm{trust}_z(s) = \min_{\mathcal{I} \in \{0,1\}^{\{r_s,m_s,e_s\}}} \max\{\tau_z(\mathcal{I}), 1 - \pi(\mathcal{I})\}, \tag{8}$$

$$\mathrm{distrust}_z(s) = \min_{\mathcal{I} \in \{0,1\}^{\{r_s,m_s,e_s\}}} \max\{\delta_z(\mathcal{I}), 1 - \pi(\mathcal{I})\}. \tag{9}$$

[4] For the the sake of readability, we restrict the interpretations $\mathcal{I}$ as if the language were built on atomic propositions r_s, m_s, and e_s only.

Besides, we can also have uncertainty about the configuration of the source's goals. The overall trust/distrust of an agent in a source s depends then on *(i)* its judgment about a source defined by τ_z/δ_z, *(ii)* the uncertainty about the source's real nature, and *(iii)* the uncertainty about the source's goals. We thus define these trust and distrust values based on goals and nature as follows:

$$\mathrm{trust}(s) = \min_{z \in \{a,b,c,d\}} \max\{\mathrm{trust}_z(s), 1 - \pi(z)\}, \tag{10}$$

$$\mathrm{distrust}(s) = \min_{z \in \{a,b,c,d\}} \max\{\mathrm{distrust}_z(s), 1 - \pi(z)\}. \tag{11}$$

Proposition 2. *If the two functions $\tau_z(\mathcal{I})$ and $\delta_z(\mathcal{I})$ are such that, for all z and $\mathcal{I}$, $\tau_z(\mathcal{I}) > 0 \Rightarrow \delta_z(\mathcal{I}) = 0$ and $\delta_z(\mathcal{I}) > 0 \Rightarrow \tau_z(\mathcal{I}) = 0$, then, for all source s, $\mathrm{trust}(s)$ and $\mathrm{distrust}(s)$ satisfy the bipolar conditions of Equations 6 and 7.*

Proof: Since π is normalized, $\exists z_0, \mathcal{I}_0$ such that $\pi(z_0) = \pi(\mathcal{I}_0) = 1$. Then, $\mathrm{trust}(s) > 0 \Rightarrow \forall z \ \max\{\mathrm{trust}_z(s), 1 - \pi(z)\} > 0 \Rightarrow \mathrm{trust}_{z_0}(s) > 0 \Rightarrow \tau_{z_0}(\mathcal{I}_0) > 0 \Rightarrow \delta_{z_0}(\mathcal{I}_0) = 0$ and $1 - \pi(\mathcal{I}_0) = 0 \Rightarrow \max\{\delta_{z_0}(\mathcal{I}_0), 1 - \pi(\mathcal{I}_0)\} = 0 \Rightarrow \mathrm{distrust}(s) = 0$. A similar reasoning proves that $\mathrm{distrust}(s) > 0 \Rightarrow \mathrm{trust}(s) = 0$. □

A corollary of this proposition is that, for all source s, $\mathrm{trust}(s) + \mathrm{distrust}(s) \leq 1$. We can notice that, in case of complete ignorance, $\mathrm{trust}(s) = \mathrm{distrust}(s) = 0$.

Definition 5. *(Trustworthiness order relation) Let s_1 and s_2 be two information sources. We consider that s_1 is less trustworthy than s_2, $s_1 \preceq s_2$, if and only if $\mathrm{trust}(s_1) \leq \mathrm{trust}(s_2)$ and $\mathrm{distrust}(s_1) \geq \mathrm{distrust}(s_2)$.*

Proposition 3. *(Total order) The relation $\preceq$ is a total order.*

Proof: The thesis is a direct consequence of Equations 6 and 7. □

4.5 Credibility and Competence

Let D be the set of domains of competence considered for the agent and the sources. The agent's competences are represented through a vector $\boldsymbol{\kappa}$, whose component κ_d represents the extent to which the agent is competent with respect to domain $d \in D$. Moreover, we suppose that the agent may have beliefs about the competences of a source. To this aim, we assume that Prop contains propositions c_d^s, meaning "source s is competent about domain d"; $\mathbf{B}(c_d^s)$ will thus be the extent to which the agent believes s is competent about d.

Definition 6. *Let $\phi \not\equiv \bot$ be new information provided by s. The extent to which the agent deems ϕ credible, given that ϕ is reported by source s, is*

$$\mathrm{cr}(\phi, s) = \max\{\mathbf{B}(\phi), \max_{d \in D} \min\{\mathrm{cr}_d(\phi, s), R(\phi, d)\}\}, \tag{12}$$

where

$$\mathrm{cr}_d(\phi, s) = \begin{cases} \min\{\mathbf{B}(c_d^s), 1 - \kappa_d, \Pi([\phi])\}, & \text{if } \kappa_d > \mathbf{B}(c_d^s); \\ \mathbf{B}(c_d^s), & \text{otherwise}. \end{cases} \tag{13}$$

For the sake of completeness, $\forall s, \mathrm{cr}(\bot, s) = 0$.

Equation 12 may be paraphrased as "ϕ being reported by s is credible if there exists a domain to which ϕ is related and with respect to which it is credible". Taking a $\max$ with $\mathbf{B}(\phi)$ accounts for the fact that the extent to which something is believable cannot be less than it is already believed, no matter which source is reporting it, since credibility is the quality of being believable. Besides, a message is believable if we deem it possible. Formally, for all formula ϕ provided by a source s we have $\mathbf{B}(\phi) \leq \mathrm{cr}(\phi, s) \leq \mathit{\Pi}([\phi])$. This definition of credibility allows us to capture the notion of "competence belief" proposed in [21], going even further by using both the receiving agent's own competence and the source's expected competences to assess the credibility of the information item.

Equation 13 involves two cases:

- if the agent is not more competent about d than the source is ($\kappa_d \leq \mathbf{B}(c_d^s)$, second case of Equation 13), then it will not try to filter the incoming message according to its own beliefs; this is a mandatory assumption if an agent is to be capable of learning from sources it believes to be more knowledgeable than it is;
- if, however, the agent believes to be more competent about d than the source ($\kappa_d > \mathbf{B}(c_d^s)$, first case of Equation 13), information supplied by the source should be evaluated by its internal credibility; in addition, the resulting credibility of supplied information should not be greater than the competence of the source providing it, otherwise an agent scarcely competent about a domain would incur the risk of accepting acritically anything that a source just a little more competent than it about that domain would say, which is not in conflict with its (admittedly very incomplete) beliefs.

Furthermore, the first case in Equation 13 refers to an "internal" credibility of ϕ, which satisfies the following two intuitive properties:

1. if ϕ is completely relevant to d, $\mathrm{cr}_d(\phi) \leq 1 - \mathbf{B}(\neg\phi)$;
2. the more the agent's knowledge is complete on domain d (i.e., the agent is competent), the more $\mathrm{cr}_d(\phi)$ will approach its lower bound $\mathbf{B}(\phi)$ and, *vice versa*, the more the agent is ignorant about d, the more it must be keen on heeding a ϕ that does not contradict its current beliefs.

Like in [6], for example, the idea here is to capture the fact that a piece of information is accepted by the agent if and only if it is "credible" for the agent. "Our definition" of credibility is nevertheless different from the one used by Booth *et al.*. They consider the set of credible formulas as "an (explicit) part of an epistemic state, since it defines how easily an agent can accept very implausible new pieces of information". In our setting, the credibility of a piece of information represents the capability of the agent to evaluate the tenability of the piece of information with respect to its own competences and the ones of the sources. Obviously, if the agent is less competent or not competent at all with respect to a domain, this credibility degree must depend (be weighted), in a sense, by the source's competence.

Proposition 4. *Given a source s, $\mathrm{cr}(\cdot, s)$ is a normalized fuzzy measure.*

Proof: We must prove that $\mathrm{cr}(\bot, s) = 0$; $\mathrm{cr}(\top, s) = 1$; and $\forall \phi, \psi \in \mathcal{L}$, $\phi \models \psi \Rightarrow \mathrm{cr}(\phi, s) \leq \mathrm{cr}(\psi, s)$ (monotonicity). Now, $\mathrm{cr}(\bot, s) = 0$ holds by definition; $\mathrm{cr}(\top, s) =$

1 holds because $\mathrm{cr}(\top, s) = \max\{\mathbf{B}(\top), \max_{d \in D} \min\{\mathrm{cr}_d(\top, s), R(\top, d)\}\} \geq \mathbf{B}(\top) = 1$. To prove monotonicity, we observe that, for every domain $d \in D$,

- $\mathbf{B}(\phi) \leq \mathbf{B}(\psi)$, because $\mathbf{B}$ is a fuzzy measure;
- $R(\phi, d) \leq R(\psi, d)$ by Proposition 1;
- $\Pi([\phi]) \leq \Pi([\psi])$, because Π is a fuzzy measure;
- $\mathbf{B}(c_d^s)$ and κ_d do not depend on ϕ or ψ.

Therefore, since $\max\{a, b\} \leq \max\{c, d\}$ and $\min\{a, b\} \leq \min\{c, d\}$ if $a \leq c$ and $b \leq d$, $\mathrm{cr}(\phi, s) \leq \mathrm{cr}(\psi, s)$. □

Being the credibility $\mathrm{cr}(\cdot, s)$ in an information content provided by a certain source a fuzzy measure, the following two properties hold:

Proposition 5. *Given a source s, $\forall \phi, \psi \in \mathcal{L}$, $\mathrm{cr}(\phi \vee \psi, s) \geq \max(\mathrm{cr}(\psi, s), \mathrm{cr}(\psi, s))$ and $\mathrm{cr}(\phi \wedge \psi, s) \leq \min(\mathrm{cr}(\psi, s), \mathrm{cr}(\psi, s))$.*

Proof: $\forall \phi, \psi \in \mathcal{L}$, (a) $\phi \models \phi \vee \psi$, and $\psi \models \phi \vee \psi$; therefore, $\mathrm{cr}(\phi, s) \leq \mathrm{cr}(\phi \vee \psi, s)$ and $\mathrm{cr}(\psi, s) \leq \mathrm{cr}(\phi \vee \psi, s)$; (b) $\phi \wedge \psi \models \phi$ and $\phi \wedge \psi \models \psi$; therefore, $\mathrm{cr}(\phi \wedge \psi, s) \leq \mathrm{cr}(\phi, s)$ and $\mathrm{cr}(\phi \wedge \psi, s) \leq \mathrm{cr}(\psi, s)$. □

4.6 Accepting Information

The extent to which a piece of information ϕ (provided by a source s) is accepted by an agent depends on the trust and distrust computed on the basis of the source's goals and nature (Equations 10 and 11) and the credibility of ϕ for the agent (Equation 12) which depends on the competences of the agent and the sources. We may be combined these values using the minimum triangular norm, to yield the extent to which ϕ provided by s is accepted by the agent:

$$\mathrm{acc}(\phi, s) = \min\{\mathrm{cr}(\phi, s), \mathrm{trust}(s)\}. \tag{14}$$

The choice of $\min$ as the aggregation operator is motivated by the fact that an agent should accept information ϕ provided by source s to the extent to which it deems ϕ credible and s trustworthy according to its goals and nature.

Proposition 6. *$\forall s$, $\mathrm{acc}(\cdot, s)$ is a fuzzy measure. It is normalized if $\mathrm{trust}(s) = 1$.*

Proof: Since $\mathrm{acc}(\cdot, s)$ is the $\min$ of a normalized fuzzy measure and $\mathrm{trust}(s)$, which is a constant for a fixed s, $\mathrm{acc}(\cdot, s)$ is a fuzzy measure, i.e., $\mathrm{acc}(\bot, s) = 0$ and, for $\phi, \psi \in \mathcal{L}$ such that $\phi \models \psi$,

$$\mathrm{acc}(\phi, s) = \min\{\mathrm{cr}(\phi, s), \mathrm{trust}(s)\} \leq \min\{\mathrm{cr}(\psi, s), \mathrm{trust}(s)\} = \mathrm{acc}(\psi, s).$$

Finally, if $\mathrm{trust}(s) = 1$, $\mathrm{acc}(\top, s) = \min\{\mathrm{cr}(\top, s), 1\} = \min\{1, 1\} = 1$. □

A piece of information ϕ may be provided by more sources. In this case, the extent to which ϕ is accepted by the agent, $\mathrm{accepted}(\phi)$ may be defined as

$$\mathrm{accepted}(\phi) = \max_{s \in \mathrm{src}(\phi)} \{\mathrm{acc}(\phi, s)\}, \tag{15}$$

where $\mathrm{src}(\phi)$ denotes the sources of ϕ. Operators other than $\max$ might be used, e.g., operators with cumulative effects. The value $\mathrm{accepted}(\phi)$ may be used as input weight for any weighted belief revision operator, like the ones studied in [3].

Experimental results obtained by Sparks in [26] shows that when untrustworthy sources provide non-credible information, individuals are less likely to revise their initial beliefs. Our formalism also captures the fact that the initial beliefs of an agent are not revised if new information is non-credible or is provided by untrustworthy sources.

5 Conclusion

The goal of this paper is to shed some light on a few fundamental formal aspects of credibility and trust as used by humans in view of their implementation on computers. More precisely, our contribution consists in providing a model for computing the acceptance of information provided by a source taking into account both trust in the source and credibility of the message.

Our model encompasses, but is not limited to, the four "kinds" of beliefs needed by an agent before delegating a task to another agent [21], where the task is "to provide useful information". In particular,

- "competence belief" is captured by "our" definition of credibility that goes even further by using also the receiving agent's own competence to assess information provided by another agent;
- "willingness belief" and "persistence belief" are captured thanks to the concept of "source nature": we should always adopt and maintain a suspicious attitude with respect to an agent we believe to be irrational for example; and
- "motivation belief" is captured by taking into account both the goals of the agent and the ones of the source: the agent believes that a source sharing its goals has some motivation to help it.

As for future work, we plan to apply our formalism towards a cognitive view of *adversarial reasoning*, and to analyze and reason over *irrational* behavior (i.e., stupid agents are dangerous because they act irrationally).

References

1. S. Adali. *Modeling Trust Context in Networks*. Springer Briefs, 2013.
2. J. Ben-Naim and H. Prade. Evaluating trustworthiness from past performances: Interval-based approaches. In *SUM*, pages 33–46, 2008.
3. S. Benferhat, C. da Costa Pereira, and A. Tettamanzi. Syntactic computation of hybrid possibilistic conditioning under uncertain inputs. In *IJCAI 2013*, 2013.
4. S. Benferhat, D. Dubois, H. Prade, and M.-A. Williams. A practical approach to revising prioritized knowledge bases. *Studia Logica*, 70(1):105–130, 2002.
5. R. Booth, E. Fermé, S. Konieczny, and R. Pino Pérez. Credibility-limited revision operators in propositional logic. In *KR 2012*, 2012.
6. R. Booth, E. Fermé, S. Konieczny, and R. Pino Pérez. Credibility-limited improvement operators. In *ECAI 2014*, pages 123–128, 2014.

7. C. Castelfranchi and R. Falcone. Social trust: A cognitive approach. In C. Castelfranchi and Y.-H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Springer, 2001.
8. C. M. Cipolla. *The basic laws of human stupidity*. il Mulino, 2011.
9. C. da Costa Pereira and A. Tettamanzi. Belief-goal relationships in possibilistic goal generation. In *ECAI 2010*, pages 641–646, 2010.
10. J. Delgrande, D. Dubois, and J. Lang. Iterated revision as prioritized merging. In *KR*, pages 210–220, 2006.
11. A. Dragoni and P. Giorgini. Belief revision through the belief-function formalism in a multi-agent environment. In *ATAL*, pages 103–115, 1996.
12. R. Falcone, M. Piunti, M. Venanzi, and C. Castelfranchi. From manifesta to krypta: The relevance of categories for trusting others. *ACM Trans. Intell. Syst. Technol.*, 4(2):27:1–27:24, 2013.
13. D. Gabbay, G. Pigozzi, and J. Woods. Controlled revision—an algorithmic approach for belief revision. *J. Log. Comput.*, 13(1):3–22, 2003.
14. P. Krümpelmann, L. Tamargo, A. García, and M. Falappa. Forwarding credible information in multi-agent systems. In *KSEM*, pages 41–53, 2009.
15. M. Lalmas. Logical models in information retrieval: Introduction and overview. *Information Processing and Management*, 34(1):19–33, January 1998.
16. M. Mas, M. Monserrat, J. Torrens, and E. Trillas. A survey on fuzzy implication functions. *Trans. Fuz Sys.*, 15(6):1107–1121, 2007.
17. D. Harrison McKnight and Norman L. Chervany. Trust and distrust definitions: One bite at a time. In *Trust in Cyber-societies*, volume 2246 of *Lecture Notes in Computer Science*, pages 27–4. Springer, 2000.
18. F. Paglieri, C. Castelfranchi, C. da Costa Pereira, R. Falcone, A. Tettamanzi, and S. Villata. Trusting the messenger because of the message: feedback dynamics from information quality to source evaluation. *Comp. & Math. Organization Theory*, 20(2):176–194, 2014.
19. I. Pinyol and J. Sabater-Mir. Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review*, 40(1):1–25, 2013.
20. R. J. Bies R. J. Lewicki, D. J. Mcallister. Trust and distrust: New relationships and realities. *The Academy of Management Review*, 23(3):438–458, 1998.
21. S. D. Ramchurn, D. Huynh, and N. R. Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(1):1–25, 2004.
22. J.P. Robinson, P.R. Shaver, and L.S. Wrightsman. *Measures of Personality and Social Psychological Attitudes*. Measures of social psychological attitudes. Academic Press, 1991.
23. J. Sabater and C. Sierra. Regret: Reputation in gregarious societies. In *Proceedings of the Fifth International Conference on Autonomous Agents*, AGENTS '01, pages 194–195, 2001.
24. G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, 1976.
25. C. Sierra and J. Debenham. An information-based model for trust. In *AAMAS*, pages 497–504. ACM, 2005.
26. J. R. Sparks and D. N. Rapp. Unreliable and anomalous: How the credibility of data affects belief revision. In *Annual Conference of the Cognitive Science Society*, pages 741–746, 2011.
27. L. Tamargo, A. García, M. Falappa, and G. Simari. Modeling knowledge dynamics in multi-agent systems based on informants. *Knowledge Eng. Review*, 27(1):87–114, 2012.
28. L. Teacy, J. Patel, N. Jennings, and M. Luck. Travos: Trust and reputation in the context of inaccurate information sources. *JAAMAS*, 12:2006, 2006.
29. L. Ughetto and V. Claveau. Different interpretations of fuzzy gradual-inclusion-based IR models. In *EUSFLAT*, pages 431–438, 2011.
30. L. Ughetto, G. Pasi, V. Claveau, O. Pivert, and P. Bosc. Implication in information retrieval systems. In *RIAO*, pages 61–64, 2010.
31. E. Ullmann-Margalit. Trust out of distrust. *J. Phil.*, 99(10):532–548, 2002.
32. L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.